DOCUMENT RESUME

ED 384 655 TM 023 948

AUTHOR Mislevy, Robert J.
TITLE A Framework for Studying Differences between Multiple-Choice and Free-Response Test Items.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-91-36
PUB DATE May 91
NOTE 59p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Cognitive Psychology; *Competence; Elementary Secondary Education; *Inferences; *Multiple Choice Tests; *Networks; Test Construction; Test Reliability; *Test Theory; Test Validity
IDENTIFIERS *Free Response Test Items

ABSTRACT

This paper lays out a framework for comparing the qualities and the quantities of information about student competence provided by multiple-choice and free-response test items. After discussing the origins of multiple-choice testing and recent influences for change, the paper outlines an "inference network" approach to test theory, in which students are characterized in terms of levels of understanding of key concepts in a learning area. It then describes how to build inference networks to address questions concerning the information about various criteria conveyed by alternative response formats. The types of questions that can be addressed in this framework include those that can be studied within the framework of standard test theory. Moreover, questions can be asked about generalized kinds of reliability and validity for inferences cast in terms of recent developments in cognitive psychology. (Contains 46 references, 4 tables, and 16 figures.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 384 655

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

A FRAMEWORK FOR STUDYING DIFFERENCES BETWEEN MULTIPLE-CHOICE AND FREE-RESPONSE TEST ITEMS

Robert J. Mislevy

BEST COPY AVAILABLE

Educational Testing Service
Princeton, New Jersey
May 1991

TM023948

A Framework for Studying Differences Between Multiple-Choice and Free-Response Test Items

Robert J. Mislevy

Educational Testing Service

April, 1991

To appear in R.E. Bennett & W.C. Ward (Eds.), *Construction vs. Choice in Cognitive Measurement*, Hillsdale, NJ: Erlbaum. The work was supported by the Program Research Planning Council of Educational Testing Service. Figures 2 and 3 were reproduced with the kind permission of Dr. Steen Andreassen.

Copyright © 1991. Educational Testing Service. All rights reserved.

A Framework for Studying Differences Between Multiple-Choice and Free-Response Test Items

Abstract

This paper lays out a framework for comparing the qualities and the quantities of information about student competence provided by multiple-choice and free-response test items. After discussing the origins of multiple-choice testing and recent influences for change, the paper outlines an "inference network" approach to test theory, in which students are characterized in terms of levels of understanding of key concepts in a learning area. It then describes how to build inference networks to address questions concerning the information about various criteria conveyed by alternative response formats. The types of questions that can be addressed in this framework include those that can be studied within the framework of standard test theory. Moreover, questions can be asked about generalized kinds of reliability and validity for inferences cast in terms of recent developments in cognitive psychology.

Introduction

Ever since Robert M. Yerkes tested a million World War I recruits with his Army Alpha Intelligence Test, multiple-choice items have dominated educational selection, placement, and assessment applications. Occasional criticism has marked their reign, from observers including no less than Banesh Hoffman, Ralph Nader, and (Educational Testing Service's own!) Norman Frederiksen. But the character of today's debates strikes at the very heart of the enterprise: The view of human ability that spawned multiple-choice tests no longer holds universal currency among psychologists and educators. The ascendent view originates from a perspective more attuned to instruction than to selection or prediction. Learners increase their competence not simply by accumulating new facts and skills at rates determined by relatively immutable "aptitudes," but by reconfiguring knowledge structures, by automating procedures and chunking information to reduce memory loads, and by developing strategies and models that tell them when and how facts and skills are relevant.

Tests can be described as mere tools to gather information, in order to guide educational decisions. But an educational decision-making framework cannot be conceived except around a view of human ability, which suggests educational options, determines what information is relevant, and specifies how an implementation is to be evaluated. The pertinent questions about multiple-choice tests now are whether, when, and how these tools, developed and validated within the old paradigm, can serve useful roles within the new paradigm.

This paper discusses an analytical framework to evaluate the contribution of different types of test items. It encompasses views of ability from the new and the old paradigms. When applied with the old, it expresses the same questions that traditional test theoretic investigations have asked. When applied with the new, it provides machinery to investigate analogous questions—about efficiency and reliability, for example. We suggest

how "inference networks" can be used to model the information from various types of observations, from traditional multiple-choice items to extended performances or portfolios, in terms of skills, strategies, and states of understanding. An application in the setting of medical diagnosis introduces the ideas, and an example based on Robert Siegler's (1981) balance beam tasks illustrates their application in a simple educational setting.

This analytic framework does not deal directly with an important issue in the debate over construction versus choice, namely, the feedback effects on instruction from the content and modality of testing (see N. Frederiksen, 1984); we focus primarily on what is learned about an individual from observations. The central role of feedback in the "why's" and "when's" of construction versus choice, however, is addressed in the background section.

Background

The Origins of Multiple-Choice Tests

The initial enthusiasm that greeted multiple-choice testing can be attributed to its neat conformance to both the psychology and the economics of the post World War I period. The psychology was that of the IQ. To the degree that a test was reliable, ranking students in accordance with overall proficiency was thought to reflect the cardinal characteristic of human intelligence. The economic character of the decisions that shaped the evolution of testing was nearly universal in education at the beginning of this century, and it dominates practice yet today: Educators were confronted with selection or placement decisions for large numbers of students, and resources limited the information they could gather about each student, constrained the number of options they could offer, and precluded much tailoring of programs to individual students once a decision was made (Glaser, 1981).

Exposing a diverse group of students to a uniform educational treatment typically produces a distribution of outcomes, as measured by a test score at the end of the program

(Bloom, 1976). Decision-makers sought to identify which students would be likely to succeed in a limited number of programs. An individual's degree of success depends on how his or her unique skills, knowledge, and interests match up with the equally multifaceted requirements of a program. The Army Alpha demonstrated that at costs substantially lower than personal interviews or performance samples, responses to multiple-choice test items could provide information about certain aspects of this matchup. What is necessary in this approach is that each item tap some of the skills required for success. Even though a single item might require only a few of the relevant skills and offer little information in its own right, a tendency to provide correct answers over a large number of items supports some degree of prediction of success along this dimension (Green, 1978). A typical multiple-choice test would be used to guide infrequent, broadly cast, decisions, such as college admittance, course placement, or military occupational assignment.

Stasis Assumptions

The multiple-choice educational testing enterprise evolved under two tacit assumptions of stasis:

1. *A characteristic of examinees that is stable over time is being measured.* In the Army Alpha and many educational tests, that characteristic is posited to be "intelligence," calibrated in units of IQ. "Intelligence" gave way to "aptitude." This assumption is consonant with infrequent decision-making; the measure should not be easily affected by short-term instruction (e.g., the SAT should not be "coachable.")
2. *The characteristics of the system affecting the examinee before the measurement is made are unaffected by the measurement process.* Focusing on the educational setting, it is assumed that the instruction prospective examinees receive and what they learn does not depend on the way their performance will be measured. It is

easy to believe this assumption is satisfied when the examiners believe, as did Yerkes, that tests are measuring examinees' "innate intelligence."

The historic cost/benefits justifications for multiple-choice testing are valid to the extent that these assumptions are met in a given testing context. Current interest in alternative modes of assessment, including constructed response formats, is spurred by the realization that these assumptions are *not* well satisfied in all educational testing problems. New types of decisions, new types of costs, and new types of benefits enter into the equation. The results in a particular context may favor multiple-choice, construction, or some mix. It is clear, however, that even when multiple-choice is favored, it must be justified in terms of the new paradigm rather than the old, a point to which we return below.

Dimensions of Change

Figure 1 offers a framework for discussing current interest in alternative modes of testing. The upper left cell represents traditional choice-type observations, used to infer overall proficiency in a specified domain. That is, it is deemed appropriate in the decision context to treat as identical all examinees with the same overall proficiency estimate, be it total score, formula score, or item response theory (IRT) ability estimate.

[Figure 1]

A first dimension of change arises from the realization that educational testing can affect educational practice. Norman Frederiksen (1984) cogently illustrates this point. During World War II, he and his coworkers encountered a gunnery course that consisted almost exclusively of bookwork and ended with a paper and pencil final exam. Frederiksen and his colleagues introduced a performance final requiring actual setup and operation of the equipment. Within a few sessions, the instructors had of their own volition introduced equipment into the course itself. Training now consisted of a balance of bookwork and performance, and trainees' final examination performance translated into

better performance in the job. Contemporary examples include the New York Regents' physics examination, where what the Regents test the students study, and high stakes assessments in schools, where rewards for high scores on specific tests encourage, at best, emphasis of those skills, and at worst, coaching of the specific items on the test.

Once it is accepted that tests do influence instruction, attention must turn to the character of that influence. A test has "systemic validity," in J. Frederiksen and Collins' (1989) terminology, if the behaviors it encourages on the part of administrators, teachers, and students enforce the learning of desired skills. Administering a final essay test can be preferable to a final multiple choice test with a higher reliability coefficient if, in preparing, students write more essays—*even if the final scores on the essay and multiple-choice tests are highly correlated* within a population of students at any given point in time. This realization spurs activity in the cell in the lower left. Students are still modeled in terms of overall proficiency, but the tasks are designed so as to better mirror the behaviors and the skills of the intended instruction. Steering instruction in positive directions constitutes a "consequential basis" for validating the use of a test (Messick, 1989).

A second dimension of change challenges the stasis of the examinee—or, more accurately, focuses attention on situations in which relatively short term change is the intended outcome of the exercise, rather than a nuisance effect. Rather than seeking long-term, stable characteristics that are immune to change, a test in this context is meant to provide information about characteristics of an examinee that are ripe for change. The problem of interest is one of diagnosis or optimal assignment to instruction; the decision is viewed as shorter term; the options are cast not in terms of level of persistent proficiency but of architecture of current proficiency. Examples would include the examinee's level of understanding of phenomena in a domain of gears and pulleys problems (Hegarty, Just, & Morrison, 1988), the mental model a student is employing for series and parallel electrical

circuit problems (Gentner & Gentner, 1983), and approaches to solving addition problems with mixed numbers (Tatsuoka, 1989).

The column on the right in Figure 1 thus concerns applications where the inference depends on the architecture of the examinee's proficiency. Information can be obtained either from choice items or from constructed responses, as will be illustrated in a following section. These possibilities are represented by the upper and lower cells respectively. The present paper focuses on comparisons of information about examinees from upper and lower cells, *within columns*. That is, the information from choice and construction items are compared given that one has already specified how to model examinees—according to overall proficiency or architecture of proficiency. It is clear from the preceding discussion that this logically precedent choice depends on the nature of the decision to be made about the student.

An Analytic Framework

In this paper, a *student model* is a caricature of a student in terms of parameters that capture distinctions that might exist among real students. A simple student model from the psychometric tradition posits a single variable—overall proficiency—that expresses all differences among students that are presumed to be relevant to the task at hand. A student model inspired by cognitive psychology might characterize a student in terms of the number of concepts and the nature of links among them. Marshall (1989; in press), for example, describes students in terms of aspects of their acquisition of schemata for arithmetic word problems. Model-based test theory consists of techniques for drawing inferences through student models, when the model for any given student cannot be specified with certainty but must be inferred imperfectly from observations.

A catalogue of such techniques has been developed over the past century for student models based on overall proficiency, from choice-type observations—the upper-left cell in Figure 1. Our interest lies in analogous techniques for the remaining cells. What is

required in a given application is a specification of the universe of potential student models and a way of connecting them to observations. Similar problems in such diverse areas as forecasting, troubleshooting, medical diagnosis, and animal husbandry have spurred research into *inference networks* (Lauritzen & Spiegelhalter, 1988; Pearl, 1988), or formal statistical frameworks for reasoning about interdependent variables in the presence of uncertainty. The next section introduces some of the key concepts in inference networks through a medical example. Following that, the ideas are related to student modeling in educational testing problems.

An Example from Medical Diagnosis

MUNIN is an inference network that organizes knowledge in the domain of electromyography—the relationships among nerves and muscles. Its function is to diagnose nerve/muscle disease states. The interested reader is referred to Andreassen, Woldbye, Falck, and Andersen (1987) for a more comprehensive description. The prototype discussed in that presentation and used for our illustration concerns a single arm muscle, with concepts represented by twenty-five nodes and their interactions represented by causal links. The ESPRIT team has generalized the application to address clusters of interrelated muscles in a network containing over a thousand nodes. A graphic representation of the simpler network appears in Figure 2.

[Figure 2]

The rightmost column of nodes in Figure 2 concerns outcomes of potentially observable variables, such as symptoms or test results. The middle layers are "pathophysiological states," or syndromes. These drive the probabilities of observations. The leftmost layer is the underlying disease state, including three possible diseases in various stages, no disease, or "Other"—a condition not built into the system. These states drive the probabilities of syndromes. It is assumed that a patient's true state can be adequately characterized by values of these disease and syndrome states. Paths indicate

conditional probability relationships, which are to be determined logically, subjectively, purely empirically, or through model-based statistical estimation. Note that the probability distribution of a given observable will depend on some syndromes, but not others. The lack of a path signifies conditional independence. Note also that a given test result can be caused by different disease combinations.

As a patient enters the clinic, the diagnostician's state of knowledge about him is expressed by population base rates. This is depicted in Figure 2 by bars that represent the base probabilities of disease and syndrome states. Base rates of observable test results are similarly shown. Tests are carried out, one at a time or in clusters, and with each result the probabilities of disease states are updated. The expectations of tests not yet given are calculated, and it can be determined which test will be most informative in identifying the disease state. Knowledge is thus accumulated in stages, with each successive test selected optimally in light of knowledge at that point in time. Figure 3 illustrates the state of knowledge after a number of electromyographic test results have been observed. Observable nodes with results now known are depicted with shaded bars representing observed values. For them, knowledge is perfect. The implications of these results have been propagated leftward to syndromes and disease states, as shown by distributions that differ from the base rates in Figure 2. These values guide the decision to test further or initiate a treatment. Finally, updated beliefs about disease states have been propagated back toward the right to update expectations about the likely outcomes of tests not yet administered. These expectations, and the potential they hold for further updating knowledge about the disease states, guide the selection of further tests.

[Figure 3]

The MUNIN example described here just concerns diagnosis; the only observable nodes are tests and symptoms. The ideas can be extended to additional types of nodes. One type would be *prognostic nodes*. Probabilities would depend on underlying disease

states. A network with diagnostic tests and prognostic assessments would draw inferences from current health indicators to likely outcomes, such as probability of survival after five years. Prognosis nodes could be potentially observable, such as whether a particular symptom will be present, or unobservable but inferable stochastically from future observables, such as the disease state that drives probabilities of new symptoms. Another type that could be introduced would be *treatment nodes*. The value of a treatment node, like an observable, would be determined with certainty when the treatment is initiated. Before this time, however, "what if" questions would be examined to explore the current projections of treatment outcomes. A prognostic node would then be affected by both disease nodes and treatment nodes; conditional probabilities of future states would depend on the current assessment of the disease state, and expected results under different treatment options. At any current state of diagnostic testing, the investigator could examine the expected results of alternative treatment options. Testing would be terminated when the additional information of subsequent tests would not provide sufficient improvement of expected treatment outcomes. For example, there may yet be several competing disease states, but if the treatment is identical in all cases, additional testing would not be warranted. (See Andreassen, Jensen, & Olesen, 1990, for a hypothetical network that encompasses diagnosis, disease identification, prognosis, and treatment selection.)

Modeling Student Understanding

To see how the ideas underlying MUNIN apply to the educational setting, consider the analogy drawn in Table 1. In collaboration with colleagues both within ETS and elsewhere, I am beginning to pursue a particular approach to student modeling based on this perspective (Mislevy, Yamamoto, & Anacker, in press). In one sense it is a natural extension of traditional psychometrics: students are described in terms of unobservable parameters, whose values, if known with certainty, would serve as the foundation for decision-making; observational settings (e.g., tests, performance observations, portfolios)

are devised that provide information about what these parameters might be; and statistical machinery is developed to guide decision-making in the face of the uncertainty engendered by ascertaining the values of only observable variables rather than parameters.

[Table 1]

Construction of an analytic framework for a specific application begins with a definition of a universe of student models. This "supermodel" is indexed by parameters that signify distinctions among states of students' understanding. Symbolically, we shall refer to the (possibly vector-valued) parameter of the student-model as η. Parameters can be qualitative and quantitative, and qualitative parameters can be unordered, partially ordered, or completely ordered. A supermodel can contain any mixture of these types. Their nature is derived from the structure and the psychology of the learning area, the idea being to capture the essential nature of key distinctions among students. A particular set of values of the parameters of the supermodel specifies a particular student model, or one particular state among the universe of possible states.

Any application poses a modeling problem, an item construction problem, and an inference problem. The following sections discuss each in turn.

The **modeling** problem is to delineate the states or levels of understanding in a learning domain. In meaningful applications this map would be expected to include several distinct strands, as understanding develops in a number of key concepts, and it might address the connectivity among the key concepts. Symbolically, this substep defines the *nature* of η and the *structure* of $p(\mathbf{x}|\eta)$, where $\mathbf{x}$ represents observations. Obviously any model will be a gross simplification of the reality of cognition. The objective is to capture differences among students that are important to the job at hand. As Greeno (1976) points out, "It may not be critical to distinguish between models differing in processing details if the details lack important implications for quality of student performance in instructional situations, or the ability of students to progress to further stages of knowledge and

understanding." For the kinds of selection decisions that spawned traditional tests, it may indeed suffice to model students solely in terms of overall proficiency. Such applications fall in the left column of Figure 1.

As useful as standard tests and standard test theory have proven in large-scale evaluation, selection, and placement problems, their focus on *who* is competent and *how many* items they can answer falls short when the goal is to improve individuals' competencies. Glaser, Lesgold, and Lajoie (1987) point out that tests can predict failure without an understanding of what causes success, but intervening to prevent failure and enhance competence requires deeper understanding. The past decade has witnessed considerable progress toward the requisite understanding. Psychological research has moved away from the traditional laboratory studies of simple (even random!) tasks, to tasks that better approximate the meaningful learning and problem-solving activities that engage people in real life. Studies comparing the ways experts differ from novices in applied problem-solving in domains such as physics, writing, and medical diagnosis (e.g., Chi, Feltovich & Glaser, 1981) reveal the central importance of knowledge structures—networks of concepts and interconnections among them—that impart meaning to patterns in what one observes and how one chooses to act. The process of learning is to a large degree expanding these structures and, importantly, *reconfiguring them* to incorporate new and qualitatively different connections as the level of understanding deepens. Educational psychologists have begun to put these findings to work in designing both instruction and tests (e.g., Glaser et al., 1987; Greeno, 1976; Marshall, 1985, in press). Again in the words of Glaser, Lesgold, and Lajoie (1987),

> "Achievement testing as we have defined it is a method of indexing stages of competence through indicators of the level of development of knowledge, skill, and cognitive process. These indicators display stages of performance that have been attained and on which further learning can proceed. They also show forms of error and misconceptions in knowledge that result in

inefficient and incomplete knowledge and skill, and that need instructional attention." (p.81)

Tests built to support such inferences lie in the rightmost column of Figure 1.

Research relevant to this approach has been carried out in a variety of fields, including cognitive psychology, the psychology of mathematics and science education, artificial intelligence (AI) work on student modeling, test theory, and statistical inference. Cognitive scientists have suggested general structures such as "frames" or "schemata" that can serve as a basis for modeling understanding (e.g., Minsky, 1975; Rumelhart, 1980), and have begun to devise tasks that probe their features (e.g., Marshall, 1989, in press). Researchers interested in the psychology of learning in subject areas such as proportional reasoning have focused on identifying key concepts, studying how they are typically acquired (e.g., in mechanics, Clement, 1982; in ratio and proportional reasoning, Karplus, Pulos, & Stage, 1983), and constructing observational settings that allow one to infer students' understanding (e.g., van den Heuvel, 1990; McDermott, 1984).

Models that focus on patterns other than overall proficiency, and which constitute rudiments for student models more consonant with the results of educational and cognitive psychology, have begun to appear in the test theory literature. Examples include the following:

- Mislevy and Verhelst's (1990) *mixture models* for item responses when different examinees follow different solution strategies or use alternative mental models.
- Falmagne's (1989) and Haertel's (1984) latent class models for *Binary Skills*. A learner is characterized as possessing or not possessing each of a number of specified skills; a task is characterized by the subset of these its solution requires. Response probabilities are driven by the matchup between the skills he or she possesses and the skills a task demands. Also see Paulson (1986) for an alternative use of latent class modelling in cognitive assessment.

- Embretson's (1985) *multicomponent models* for integrating item construction and inference within a unified cognitive model. The conditional probabilities of solution steps given a multifaceted student model are given by IRT-like statistical structures.
- Tatsuoka's (1989) *Rule space* analysis. Tatsuoka uses a generalization of IRT methodology to define a metric for classifying examinees based on likely patterns of item response given patterns of knowledge and strategies.
- Masters and Mislevy's (in press) and Wilson's (1989a) use of the *Partial Credit* rating scale model to characterize levels of understanding, as evidenced by the nature or approach of a performance rather than its correctness.
- Wilson's (1989b) *Saltus* model for characterizing stages of conceptual development. Item responses are assumed to follow an IRT model within stages, but the characteristics of items are allowed to differ across stages.
- Yamamoto's (1987) *Hybrid* model for dichotomous responses. The *Hybrid* model characterizes an examinee as belonging either to one of a number of classes associated with specified states of understanding, or in a catch-all IRT class. Examinees in this catch-all class are characterized merely as to overall proficiency; their response patterns are not strongly associated with the states that are built into the model.

The **item construction** problem is to devise situations in which students who differ in the parameter space are likely to behave in observably different ways. The conditional probabilities of behavior of different types given the unobservable state of the student are the *values* of $p(x|\eta)$, which may in turn be modeled in terms of another set of parameters, say β. The $p(x|\eta)$ values provide the basis for inferring back about the student state. For measuring overall proficiency, $p(x|\eta)$ might take the form of an IRT model, with item parameters β; examinees with high proficiency should be more likely than those with low proficiency to provide correct answers. As an example of the architecture of

proficiency, Gentner and Gentner (1983) discuss how different parallel and series combinations of resistors and batteries prove differentially difficult for students using "water flow" as opposed to "teeming crowds" analogies to solve electrical circuit problems; items would be devised to distinguish between students using one analogy or the other, or neither or both.

Whatever the character of the student model, an element in $\mathbf{x}$ could contain a right or wrong answer to a multiple-choice test item, but it could instead be the problem-solving approach regardless of whether the answer is right or wrong, the quickness of responding, a characteristic of a think-aloud protocol, or an expert's evaluation of a particular aspect of the performance. These distinctions determine whether one is operating in the top row or the bottom row of Figure 1.

Specifying a universe of student models selects a column of Figure 1. This step depends on the nature of the inference or decision to be made. The question of choice versus construction is well defined when, for a given student model, observations of both types can be gathered. The decision of the mix to be observed can depend, wholly or partly, on the amount of information conveyed by alternative observations about the unobservable parameters in the student model. The effectiveness of an item is reflected in differences in conditional probabilities associated with different parameter configurations, so an item may be very useful in distinguishing among some aspects of student models but useless for distinguishing among others.

The **inference** problem is to reason from observations to student models. The model-building and item construction steps provide η and $p(\mathbf{x}|\eta)$. Let $p(\eta)$ represent expectations about η in a population of interest—possibly non-informative, possibly based on expert opinion or previous analyses. Bayes' theorem can be employed to draw inferences about η given $\mathbf{x}$ via $p(\eta|\mathbf{x}) \propto p(\mathbf{x}|\eta)\, p(\eta)$. Thus $p(\eta|\mathbf{x})$ characterizes belief about a particular student's model after having observed a sample of the student's behavior.

Practical problems include characterizing what is known about β so as to determine $p(x|\eta)$, carrying out the computations involved in determining $p(\eta|x)$, and, in some applications, developing strategies for efficient sequential gathering of observations. The ESPRIT team that developed MUNIN has developed the inference network shell HUGIN (Andersen, Jensen, Olesen, & Jensen, 1989) to carry out calculations of this type, using the computational advances introduced by Lauritzen & Spiegelhalter (1988).

As mentioned above in the MUNIN example, an inference network can be extended with prognostic nodes and treatment nodes. In educational selection, a prognostic node might be a rating of success in a training course. The goal would be to gather information about an examinee until it could be predicted with sufficient accuracy whether the rating would be above or below a cutpoint. In instructional assignment, the same prognostic node could be used but predictions would depend on instructional options as well. Now the goal would be to determine to a sufficient degree of accuracy which instructional option gives the highest expectation of success. For example, the determination may depend on identifying the mental model a student is employing, in order to explicate the limitations of that model and introduce complementary models.

Application to Overall Proficiency

IRT models are special instances of inference networks, with the form shown as Figure 4. There is one unobservable node in a basic IRT model, the overall proficiency, often denoted θ in the IRT literature but denoted η here for consistency. There is one observable node for each test item, x_j for j=1,...,n. A link runs from η to each x_j, symbolizing the conditional probability distribution of the potential responses. If the only potential responses are right or wrong, this is just a probability of a correct response at each η value; that is, $P(x_j=1|\eta)$. The lack of links among items indicates the assumption of local, or conditional, independence of item responses given η. Most applications of IRT with multiple-choice items attend only to overall correctness, although models are also

available for partial-credit scoring of choices, where the alternatives are distinguished and the conditional probabilities of each are modeled as functions of overall proficiency (e.g., Bock, 1972; Thissen & Steinberg, 1984; Samejima, 1979; Sympson, 1983).

[Figure 4]

Item response theory models can also be employed with constructed response items. What is required is that a score of some sort be assigned to a response, bringing it into the same framework as the choice items discussed immediately above—that is, responses are mapped onto a scoring scale, and scores are mapped to η via an IRT model. The mapping of responses to item scores may be done by a human judge or mechanically in accordance with rules (Bennett, in press). An additional option is for the score to be a continuous real-valued number; Samejima (1973) provides an IRT model of this type. Figure 4 still represents the inference network.

Consider an application in which both choice and constructed responses can be garnered. How can the value of their information be compared? We consider four possibilities, distinguished by whether one wishes to allow for the possibility that "overall proficiency" has different meanings for choice and construction items (single proficiency versus distinct proficiencies) and whether a prognostic node is included.

A Single Proficiency, No Criterion Available

Figure 4 is the appropriate inference network in this case. There is only one proficiency; probabilities of success to all items are driven by that proficiency alone, and are independent otherwise. Some of the observable nodes correspond to choice items, others correspond to construction items. A single IRT model is fit to data in which samples of examinees have been administered overlapping subsets of items of both types. Comparing information is straightforward in this case. Start with the "nothing known about an individual" state, analogous to the new patient in the MUNIN example. How much is the posterior distribution for his or her η sharpened by ascertaining a subset of

choice items? How much by a subset of construction items? How much by various mixes? How much for subsets of items of the two types that take the same amount of time to administer? Do the answers vary for examinees at low, medium, or high proficiencies, as determined by entering typical responses at these various η levels? These are analyses of the accuracy with which different tests distinguish among student models within the "single overall proficiency" supermodel—in traditional terminology, the reliabilities of different tests that could be constructed from the full pool of items of all types.[1]

A Single Proficiency, Criterion Available

Extending the analytic framework of the preceding paragraph to predictive potential is accomplished by including one or more prognostic or criterion nodes, as illustrated in Figure 5. Having assumed that the same proficiency drives probabilities for both choice and construction items means that an advantage in accuracy for η translates directly into an advantage in accuracy for the criterion. What is new is the ability to evaluate information in terms of predictive power ("predictive validity") rather than reliability.

[Figure 5]

Distinct Proficiencies, No Criterion Available

Empirical evidence suggests distinctions in at least some tests, in at least some examinee populations, between overall proficiency on choice items and overall proficiency on construction items. In the College Board's Advanced Placement (AP) History examinations, for example, girls appear to enjoy an advantage over boys on the essays compared to multiple choice, while in AP in Calculus, the boys do relatively better in the open-ended problems (Mazzeo, Schmidt, & Bleistein, in press). Figure 6 allows for the possibility of two distinct proficiencies. Choice items are driven by one, with conditional probabilities expressed perhaps through an IRT model, and construction items are driven by another. The link between the two proficiencies allows for the (possibly high) relationship between them.

[Figure 6]

The link between the two proficiencies allows belief about one proficiency to be updated by information from items of the other type. This suits an application in which one proficiency—say, the one driving construction items—is ultimately of interest, yet information about it can be obtained indirectly from the other—say, from choice items. If it is sufficiently easier or less expensive to secure, such indirect information can update belief about the node of interest more efficiently than information from nodes linked to it directly.

Distinct Proficiencies, Criterion Available

The nodes and links added in Figure 7 introduce a prognostic variable. Interesting possibilities occur because information can flow to the criterion variable from both proficiencies, which can differ in their strength. The argument that in a particular application, construction items are less reliable but more valid than choice items requires the complexity of Figure 7 as opposed to Figure 5. This claim requires (1) weaker conditional probability links from construction item nodes to construction proficiency than those from choice item nodes to choice proficiency, (2) a commensurately stronger link from construction proficiency to the criterion than from choice proficiency to the criterion, and (3) a relatively weak link between choice and construction proficiencies.

[Figure 7]

Application to Proficiency Architectures

This section uses similar reasoning to compare the information from choice and construction items when the latent variables are more complex than simply low-to-high proficiency. To fix ideas, we employ the balance beam example from Mislevy, Yamamoto, and Anacker (in press).

Siegler's Balance Beam Tasks

Piaget studied children's developing understanding of proportion with a variety of methods, including their explanations of balance beam problems (Inhelder & Piaget, 1958).

Robert Siegler (1981) devised a set of balance beam tasks such that patterns of response could be predicted from the stages Piaget delineated. The tasks are exemplified in Figure 8. Varying numbers of weights are placed at different locations on a balance beam; the child predicts whether the beam will tip left, tip right, or remain in balance. Piaget would posit that they will respond in accordance with their stage of understanding, the typical progression of which is outlined below. Data from the tasks are indistinguishable from standard multiple-choice test data on the surface, but there are two key distinctions:

1. What is important about examinees is not their overall probability of answering items correctly, but their (unobservable) state of understanding of the domain.
2. Children at less sophisticated levels of understanding initially get certain problems right for the wrong reasons. These items are more likely to be answered wrong at intermediate stages, as understanding deepens! They are bad items by the standards of classical test theory and IRT, because probabilities of correct response do not increase monotonically with increasing total test score. From the perspective of the developmental theory, however, not only is this reversal expected, but it is instrumental in distinguishing among children with different ways of thinking about the problems.

[Figure 8]

The usual stages through which children progress can be described in terms of successive acquisition of the rules listed below.

Rule I: If the weights on both sides are equal, it will balance. If they are not equal, the side with the heavier weight will go down. (Weight is the "dominant dimension," because children are generally aware that weight is important in the problem earlier than they realize that distance from the fulcrum, the "subordinate dimension," also matters.)

Rule II: If the weights and distances on both sides are equal, then the beam will balance. If the weights are equal but the distances are not, the side with the longer distance will go down. Otherwise, the side with the heavier weight will go down. (A child using this rule uses the subordinate dimension only when information from the dominant dimension is equivocal.)

Rule III: Same as Rule II, except that if the values of both weight and length are unequal on both sides, the child will "muddle through" (Siegler, 1981, p.6). (A child using this rule now knows that both dimensions matter, but doesn't know just how they combine. Responses will be based on a strategy such as guessing.)

Rule IV: Combine weights and lengths correctly (i.e., compare torques, or products of weights and distances).

It was thus hypothesized that each child could be classified into one of five stages—the four characterized by the rules, or an earlier "preoperational" stage in which it is not recognized that either weight or length bear any systematic relationship to the action of the beam. The classification of students is a simple example of the "architecture of proficiency," placing it in the right-hand column of Figure 1. While Piaget's interviews fall in the lower cell, Siegler's tasks fall in the upper. Table 2 shows the probabilities of correct response that would be expected from groups of children in different stages, if their responses were in complete accordance with the hypothesized rules.[2] Scanning across the rows reveals how the probability of a correct response to a given type of item does not always increase as level of understanding increases. For example, Stage II children tend to answer Conflict-Dominant items right for the wrong reason, while Stage III children, aware of a conflict, flounder.

[Table 2]

A Latent Class Model for Balance Beam Tasks

If the theory were perfect, the columns in Table 2 would give probabilities of correct response to the various types of items from children at different stages of understanding. Observing a correct response to a Subordinate item, for example, would eliminate the possibility that the child was in Stage I. But because the model is not perfect, and because children make slips and lucky guesses, any response could be observed from a child in any stage. A latent class model (Lazarsfeld, 1950) can be used to express the structure posited in Table 2 while allowing for some "noise" in real data (see Mislevy et al., in press, for details). Instead of expecting incorrect responses with probability one to Subordinate items from Stage I children, we might posit some small fraction of correct answers—p(Subordinate correct|Stage=I). Similar probabilities of "false positives" can be estimated for other cells in Table 2 containing 0's. In the same spirit, probabilities less than one, due to "false negatives," can be estimated for the cells with 1's. Inferences cannot be as strong when these uncertainties are present; a correct response to a Subordinate item still suggests that a child is probably not in Stage I, but no longer is it proof positive.

Expressing this model in the notation introduced above, η represents stage membership, $\mathbf{x}$ represents item responses, and $p(\mathbf{x}|\eta)$ are conditional probabilities of correct responses to items of the various types from children in different stages—a noisy version of Table 2. The proportions of children in a population of interest at the different stages are $p(\eta)$, and the probabilities that convey our knowledge about a child's stage after we have observed his responses are $p(\eta|\mathbf{x})$.

Siegler created a 24-task test comprised of four tasks of each type. These tasks can be considered multiple-choice, because the respondent was asked to predict for each whether the beam would tip left, tip right, or balance—only one of which would actually happen. Siegler collected data from 60 children, from age 3 up through college age, at two points in time, for a total of 120 response vectors. Mislevy et al. (in press) fit a latent class

model to these data using the HYBRIL computer program (Yamamoto, 1987), obtaining the conditional probabilities—$p(x|\eta)$—shown in Table 3, and the following vector summarizing the (estimated) population distribution of stage membership:

$$p(\eta) = (\text{Prob(Stage=0)}, \text{Prob(Stage=I)}, ..., \text{Prob(Stage=IV)})$$
$$= (.257,.227,.163,.275,.078) .$$

[Table 3]

Note that different types of items are differentially useful to distinguish among children at different levels. Equal items, for example, are best for distinguishing Stage 0 children from everyone else. Conflict-Dominant items, which would be dropped from standard tests because their probabilities of correct response do not have a strictly increasing relationship with total scores, help differentiate among children at Stages II, III, and IV.

Figure 9 depicts the state of knowledge about a child before observing any responses. Just one item of each type is shown rather than all four for simplicity. The corresponding status of an observable node (i.e., an item type) is the expectation of a correct response from a child selected at random from the population. The path from the stage-membership node to a particular observable node represents a row of Table 3.

[Figure 9]

Figure 9 represents the state of our knowledge about a child's reasoning stage and expected responses before any actual responses are observed. How does knowledge change when a response is observed? One of the children in the sample, Douglas, gave an incorrect response to his first Subordinate item. This could happen regardless of Douglas' true stage; the probabilities are obtained by subtracting the entries in the S row of Table 3 from 1.000, yielding, for Stages 0 through IV, .667, .973, .116, .019, and .057 respectively. This is the likelihood function for η induced by the observation of the response. The bulk of the evidence is for Stages 0 and I. Combining these values with the initial stage probabilities $p(\eta)$ via Bayes' theorem yields updated stage probabilities,

p(η|incorrect response to a Subordinate item): for Stages 0 through IV respectively, .41, .52, .04, .01, and .01. Expectations for items not yet administered also change. They are averages of the probabilities of correct response expected from the various stages, now weighted by the new stage membership probabilities. The state of knowledge after observing Douglas' first response is depicted in Figure 10 (see Mislevy et al., in press, for computational details; also see Macready & Dayton, 1989.)

[Figure 10]

Extending the Paradigm

The balance beam exemplar illustrates the challenge of inferring states of understanding, but it addresses development of only a single key concept. A broader view characterizes interconnections among distinct elements of understanding or lines of development. Calculating and comparing torques to solve the "conflict" problems characterizes Stage IV. But if a child at Stage IV cannot carry out the calculations reliably, his pattern of correct and incorrect responses would be hard to distinguish from that of a child in Stage III. Although the two children might answer about the same number of items correctly, the instruction appropriate for them would differ dramatically. And children at any stage of understanding of the balance beam might be able to carry out the computational operations in isolation. This section discusses a hypothetical extension to the exemplar, namely, the ability to carry out the arithmetic operations needed to calculate torques. For illustrative purposes, we simply posit a skill to carry these calculations out reliably, either possessed by a child or not.[3] The goal of the extended system is to infer both balance-beam understanding and computational skill. To make the distinctions among states of understanding in this extended domain, we introduce two new types of observations:

1. Items isolating computation, such as "Which is greater, 3x4 or 5x2?"

2. Probes for introspection about solutions to conflict items: "How did you get your answer?" These items are construction items, in contrast to the choice items asking simply for a prediction, of "tip left," "tip right," or "remain in balance."

Figure 11 offers one possible structure for this network. The "S-Correct" and "CS-Correct" nodes represent the multiple-choice responses: is the stated prediction correct? The "CS-Why" node represents the examinee's verbal explanation for a prediction for a particular Conflict-Subordinate task. To keep the diagram simple, only one balance-beam task each for a Subordinate and a Conflict-Subordinate task are illustrated. Equal and Dominant tasks would have the same paths as the Subordinate task, and Conflict-Dominant and Conflict-Equal tasks would have the same paths as the Conflict-Subordinate tasks, although the conditional probability values would generally differ.

[Figure 11]

There are three unobservable variables in the system; i.e., η has three components. The first again expresses level of understanding in the balance beam domain. The second is the ability to carry out the calculations involved in computing torques. The third concerns the integration of balance-beam understanding and calculating proficiency. Specifically, it indicates whether a child both is in Stage IV and possesses the requisite computational skills.

Before discussing the construction versus choice tradeoffs in this network, we mention in passing some conditional independence assumptions implicit in the figure. First, note that the probabilities of the pure computation items depend on the unobservable computation variable only; they are conditionally independent of level of balance beam understanding. Secondly, for children in Stages 0 through III, both the right/wrong answers and their explanations depend only on level of understanding. Because they do not realize the connection between the problems and the torque calculations, their responses

to the balance beam tasks are conditionally independent of their computational skill, even on items for which that skill is an integral component of an expert solution.

The correctness aspect of an answer has only two possibilities, right or wrong, but an explanation can fall into five categories corresponding to stages of understanding. The overt explanation is the raw observation of the constructed answer; an expert's judgement categorizes that response into one of the stage categories. Note that a Stage III child might give an explanation consistent with Stages 0, I, II, or III, but would not, by definition, give a Stage IV explanation. Theory thus posits that the conditional probability of a Stage K response from a Stage J child is zero if K>J. Conditional probabilities for K≤J might be estimated from data or based on experts' experience. The most likely explanation for an Equal task from people at Stage IV would probably be a Stage II explanation: "It balances because both the weights and distances are equal." A hypothetical example of the conditional probabilities of explanations of different levels for a Equal item are given in Table 4.

[Table 4]

The first type of comparison of information from construction and choice items occurs when a given inference can be drawn from items of either type: to what degree do responses update the nodes of interest? In this example, probabilities of stage membership can be updated by observing either choice or construction data. Their comparative values depend on the conditional probabilities $p(\eta|x)$ associated with the potential responses. Specifically, the more the probabilities from a given latent state are concentrated on a few observable states, the more "reliable" the items are. The construction items in this hypothetical example would be more reliable in this sense than the choice items because a correct prediction can emanate from a student at any level, whereas a Level K verbal explanation can only come from a student at Level K or higher.

The second basis of comparison is whether certain inferences can be drawn from one type of response but not the other. For children in Stage IV, right/wrong answers to conflict items depend on the understanding/computation integration variable, but explanations depend only on understanding. As noted above, information from choice items alone cannot determine whether poor performance on conflict items is due to Stage III reasoning or Stage IV reasoning coupled with an inability to carry out calculations reliably. This is represented in Figure 12: after a correct answer to a choice-type Subordinate item and an incorrect answer to a choice-type Conflict-Subordinate item, the state of knowledge expresses a mixture of four possibilities:

1. The student is in a low stage (0-II) and gave a correct answer to the Subordinate item without a correct rationale.
2. The student has the requisite computational skill but is in Stage III, and thus answered the Conflict-Subordinate item incorrectly because she did not have sufficient understanding.
3. The student is in Stage IV but lacks the requisite computational skill, and thus answered the Conflict-Subordinate item incorrectly because of an error when carrying the appropriate torque comparison.
4. The student is in Stage III *and* lacks the requisite computational skill.

[Figure 12]

All four of these possibilities lead to the prediction of an incorrect response to future Conflict-Subordinate items or a structurally similar proportional reasoning task such as "shadows" (Siegler, 1981). Figure 13 illustrates this by adding a prognostic node: a Conflict-Subordinate choice-item at Time 2. For an accurate prediction, or a selection decision based on this composite skill, the distinction among the four possibilities would probably *not* be important; the child would probably perform with a similar pattern, for the same *unknown* reasons. But if the distinction were important to make—as it would be if

the objective is to improve the student's chances at handling such problems—it could be accomplished by accumulating information of additional kinds. One way would be to obtain responses to open-ended or choice-type computation problems as in Figure 14, to investigate the hypothesis of calculation failure. Alternatively, one could obtain an open-ended answer to the Conflict-Subordinate item—why did you make the response you did? Figure 15 illustrates this possibility. The results of these types of items would inform whether the instruction should be computational or conceptual. Figure 16 adds instructional nodes, and illustrates the case in which information from a choice item has indicated that the problem was computational, and computational instruction increases the probability of a correct response at Time 2.

[Figures 13-16]

A Final Comment

This paper is based on the following premises:

1. Educational testing is gathering information to make educational decisions.
2. The type of decisions to be made determines the student model that is appropriate. It is assumed that the decision maker would know how to act if a particular student's model were known with certainty.
3. Student models cannot be known with certainty. At any point in time, the model for a given student is known only up to the probabilities of various alternatives.
4. Test theory consists of rules and techniques for designing observational settings to obtain information to reduce uncertainty about student models, and for carrying out inferences and making decisions in the presence of remaining uncertainty.
5. Choice and construction items are two options for obtaining information to update a student model.
6. Decisions about the type of observations to gather should be made in light of (a) the information they provide for the decision at hand, and (b) the real and potential

consequences of the method of testing for the system in which the testing takes place.

The paper's objectives are achieved to the extent that it places the debate on choice and construction in a framework wherein advocates of different approaches can ask questions that have common meanings to all participants, and that can, at least in principle, be answered.

Notes

1. As noted in the introduction, neither this analytic framework nor those that follow take into account feedback effects on the educational system, which can, in cost/benefit analyses, tip the balance in favor of tests with lower reliabilities and predictive potentials.

2. The values in Table 2 assume that whenever a child's state of understanding does not predict a particular answer, the probabilities of responding with "tip left," "tip right," and "equal" are the same. Propensities to respond one way or another will certainly exist within particular children in these stages, and they may vary systematically with stage of understanding. These probabilities could be estimated from richer data, as might be gathered from the hypothetical extension of the test described below.

3. Obviously, states of understanding for calculating and comparing torques could be developed in greater detail, and would indeed have to be if one intended to remediate skill deficits in this domain. Verifying the presence of the broadly construed skill suffices to eliminate it as a source of failure on Conflict beam items. Discovering its absence could trigger further investigation with an inference network probing the details of the composite skill's architecture.

References

Andersen, S.K., Jensen, F.V., Olesen, K.G., & Jensen, F. (1989). HUGIN: A shell for building Bayesian belief universes for expert systems [computer program]. Aalborg, Denmark: HUGIN Expert Ltd.

Andreassen, S., Woldbye, M., Falck, & Andersen, S.K. (1987). MUNIN: A causal probabilistic network for interpretation of electromyographic findings. Proceedings of the 10th International Joint Conference on Artificial Intelligence (pp. 366-372). Milan: Kaufmann.

Bennett, R.E. (in press). Toward intelligent assessment: An integration of constructed response testing, artificial intelligence, and model-based measurement. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.

Bloom, B.S. (1976). Human characteristics and school learning. New York: McGraw-Hill.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-52.

Chi, M.T.H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Clement, J. (1982). Students' preconceptions of introductory mechanics. American Journal of Physics, 50, 66-71.

Embretson, S.E. (1985). Multicomponent latent trait models for test design. In S.E. Embretson (Ed.), Test design: Developments in psychology and psychometrics. Orlando: Academic Press.

Falmagne, J-C. (1989). A latent trait model via a stochastic learning theory for a knowledge space. Psychometrika, 54, 283-303.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193-202.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18, 27-32.

Gentner, D., & Gentner, D.R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A.L. Stevens (Eds.), Mental models (pp. 99-129). Hillsdale, NJ: Erlbaum.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. American Psychologist, 36, 923-936.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J.C. Conoley, & J. Witt (Eds.), The influence of cognitive psychology on testing and measurement: The Buros-Nebraska Symposium on measurement and testing, Vol. 3 (pp. 41-85). Hillsdale, NJ: Erlbaum.

Green, B. F. (1978) In defense of measurement. American Psychologist, 33, 664-670.

Greeno, J.G. (1976). Cognitive objectives of instruction: Theory of knowledge for solving problems and answering questions. In D. Klahr (Ed.), Cognition and instruction (pp. 123-159). Hillsdale, NJ: Erlbaum.

Haertel, E.H. (1984). An application of latent class models to assessment data. Applied Psychological Measurement, 8, 333-346.

Hegarty, M., Just, M.A., & Morrison, I.R. (1988). Mental models of mechanical systems: Individual differences in qualitative and quantitative reasoning. Cognitive Psychology, 20, 191-236.

Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic.

Karplus, R., Pulos, S., & Stage, E. (1983). Proportional reasoning of early adolescents. In R.A. Lesh & M. Landau (Eds.), Acquisition of mathematics concepts and processes (pp. 45-90). Orlando, FL: Academic Press.

Lauritzen, S.L., & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society, Series B, 50, 157-224.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Clausen, Studies in social psychology in World War II, Volume 4: Measurement and prediction (pp. 362-412). Princeton, NJ: Princeton university Press.

Macready, G.B., & Dayton, C.M. (1989, March). Adaptive testing with latent class models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Marshall, S.P. (1985, December). Using schema knowledge to solve story problems. Paper presented at the Office of Naval Research Contractors' Conference, San Diego, CA.

Marshall, S.P. (1989). Generating good items for diagnostic tests. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 433-452). Hillsdale, NJ: Erlbaum.

Marshall, S.P. (in press). Assessing schema knowledge. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.

Masters, G., & Mislevy, R.J. (in press). New views of student learning: Implications for educational measurement. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.

Mazzeo, J., Schmidt, A., & Bleistein, C. (in press). Exploratory analyses of some possible causes for the discrepencies in gender differences on multiple-choice and free response sections of the Advanced Placement examinations. Princeton NJ: Educational Testing Service.

McDermott, L.C. (1984). Research on conceptual understanding in mechanics. Physics Today, 37, 24-32.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational measurement (3rd Ed.) (pp. 13-103). New York: American Council on Education/Macmillan.

Minsky, M. (1975). A framework for representing knowledge. In P.H. Winston (Ed.), The psychology of computer vision (pp. 211-277). New York: McGraw-Hill.

Mislevy, R.J., & Verhelst, N. (1990). Modeling item responses when different subjects follow different solution strategies. Psychometrika, 55, 195-215.

Mislevy, R.J., Yamamoto, K, & Anacker, S. (in press). Toward a test theory for assessing student understanding. In R.A. Lesh (Ed.), Assessing higher-level understanding in middle-school mathematics. Hillsdale, NJ: Erlbaum.

Paulsen, J.A. (1986). Latent class representation of systematic patterns in test responses. Technical Report ONR-1. Portland, OR: Psychology Department, Portland State University.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Mateo, CA: Kaufmann.

Rumelhart, D.A. (1980). Schemata: The building blocks of cognition. In R. Spiro, B. Bruce, & W. Brewer (Eds.), Theoretical issues in reading comprehension (pp. 33-58). Hillsdale, NJ: Erlbaum.

Samejima, F. (1973). Homogeneous case of the continuous response level. Psychometrika, 38, 203-219.

Samejima, F. (1979). A new family of models for the multiple- choice item. ONR Research Report 79-4. Knoxville, TN: University of Tennessee.

Siegler, R.S. (1981). Developmental sequences within and between concepts. Monograph of the Society for Research in child Development, 46.

Sympson, J.B. (1983). A new item response theory model for calibrating multiple-choice items. Paper presented at the annual meeting of the Psychometric society, Los Angeles, CA.

Tatsuoka, K.K. (1989). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto, (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 453-488). Hillsdale, NJ: Erlbaum.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. Psychometrika, 47, 201-214.

van den Heuvel, M. (1990). Realistic arithmetic/mathematics instruction and tests. In K. Gravemeijer, M. van den Heuvel, & L. Streefland (Eds.), Context free productions tests and geometry in realistic mathematics education (pp. 53-78). Utrecht, The Netherlands: Research Group for Mathematical Education and Educational computer Center, State University of Utrecht.

Wilson, M.R. (1989a). A comparison of deterministic and probabilistic approaches to measuring learning structures. Australian Journal of Education, 33, 125-138.

Wilson, M.R. (1989b). Saltus: A psychometric model of discontinuity in cognitive development. Psychological Bulletin, 105, 276-289.

Yamamoto, K. (1987). A model that combines IRT and latent class models. Unpublished doctoral dissertation, University of Illinois.

Table 1

Inference Networks in Medicine and Education

Medical Application	Educational Application
Observable symptoms, medical tests	Test items, verbal protocols, teachers' ratings of student performances
Disease states, syndromes	States or levels of understanding of key concepts, strategy choices
Interconnections based on medical theory	Interconnections based on cognitive and educational theory
Medical prognosis	Predictive distribution for criterion measures
Evaluation of potential treatment options	Expectations of student status after potential educational treatment

Table 2

Theoretical Conditional Probabilities--Expected Proportions of Correct Response

Problem type	Stage 0	Stage I	Stage II	Stage III	Stage IV
Equal	.333	1.000	1.000	1.000	1.000
Dominant	.333	1.000	1.000	1.000	1.000
Subordinate	.333	.000	1.000	1.000	1.000
Conflict-Dominant	.333	1.000	1.000	.333	1.000
Conflict-Subordinate	.333	.000	.000	.333	1.000
Conflict-Equal	.333	.000	.000	.333	1.000

Table 3

Estimated Conditional Probabilities--Expected Proportions of Correct Response

Problem type	Stage 0	Stage I	Stage II	Stage III	Stage IV
Equal	.333*	.973	.883	.981	.943
Dominant	.333*	.973	.883	.981	.943
Subordinate	.333*	.026	.883	.981	.943
Conflict-					
Dominant	.333*	.973	.883	.333*	.943
Conflict-					
Subordinate	.333*	.026	.116	.333*	.943
Conflict-Equal	.333*	.026	.116	.333*	.943

* denotes fixed value

Table 4

Conditional Probabilities of Explanations to an E Item

Respondent's Stage	Explanation				
	Stage 0	Stage I	Stage II	Stage III	Stage IV
Stage 0	1	0	0	0	0
Stage I	.20	.80	0	0	0
Stage II	.10	.10	.80	0	0
Stage III	.05	.05	.80	.20	0
Stage IV	.03	.10	.70	.02	.15

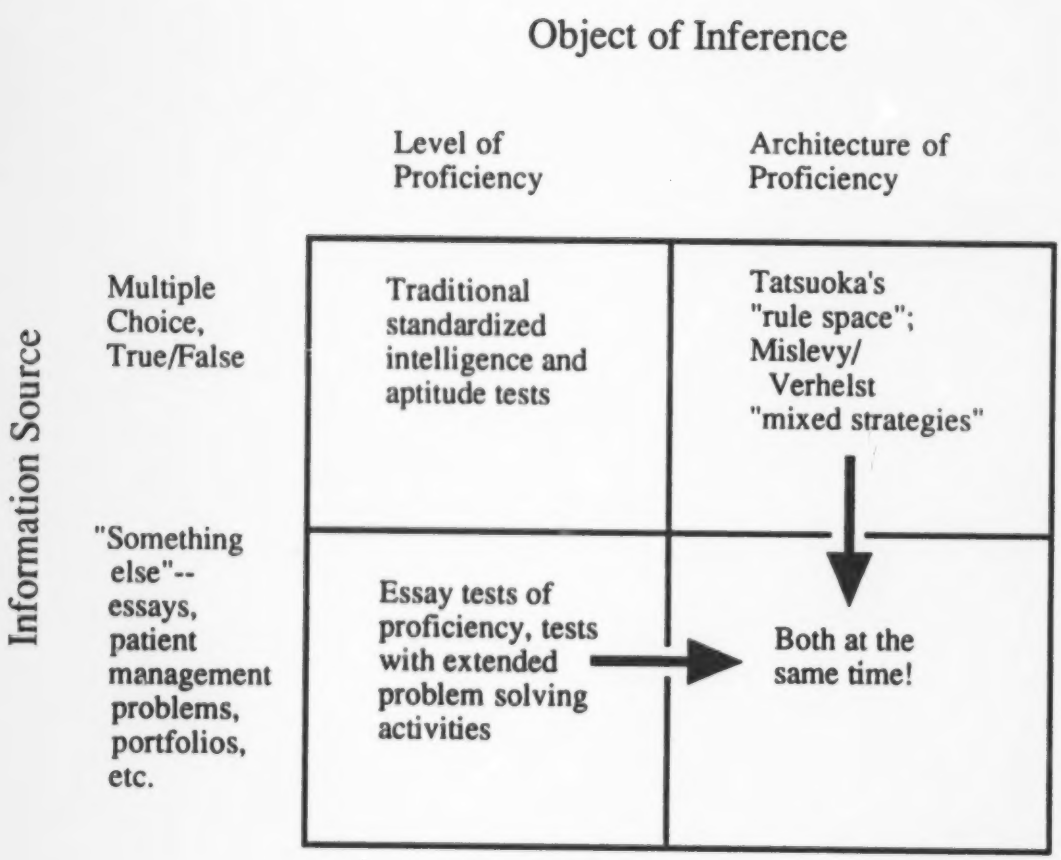

Object of Inference
Level of Proficiency
Architecture of Proficiency
Information Source
Multiple Choice, True/False
"Something else"-- essays, patient management problems, portfolios, etc.
Traditional standardized intelligence and aptitude tests
Tatsuoka's "rule space"; Mislevy/ Verhelst "mixed strategies"
Essay tests of proficiency, tests with extended problem solving activities
Both at the same time!

Figure 1

Modes and Objects of Educational Testing

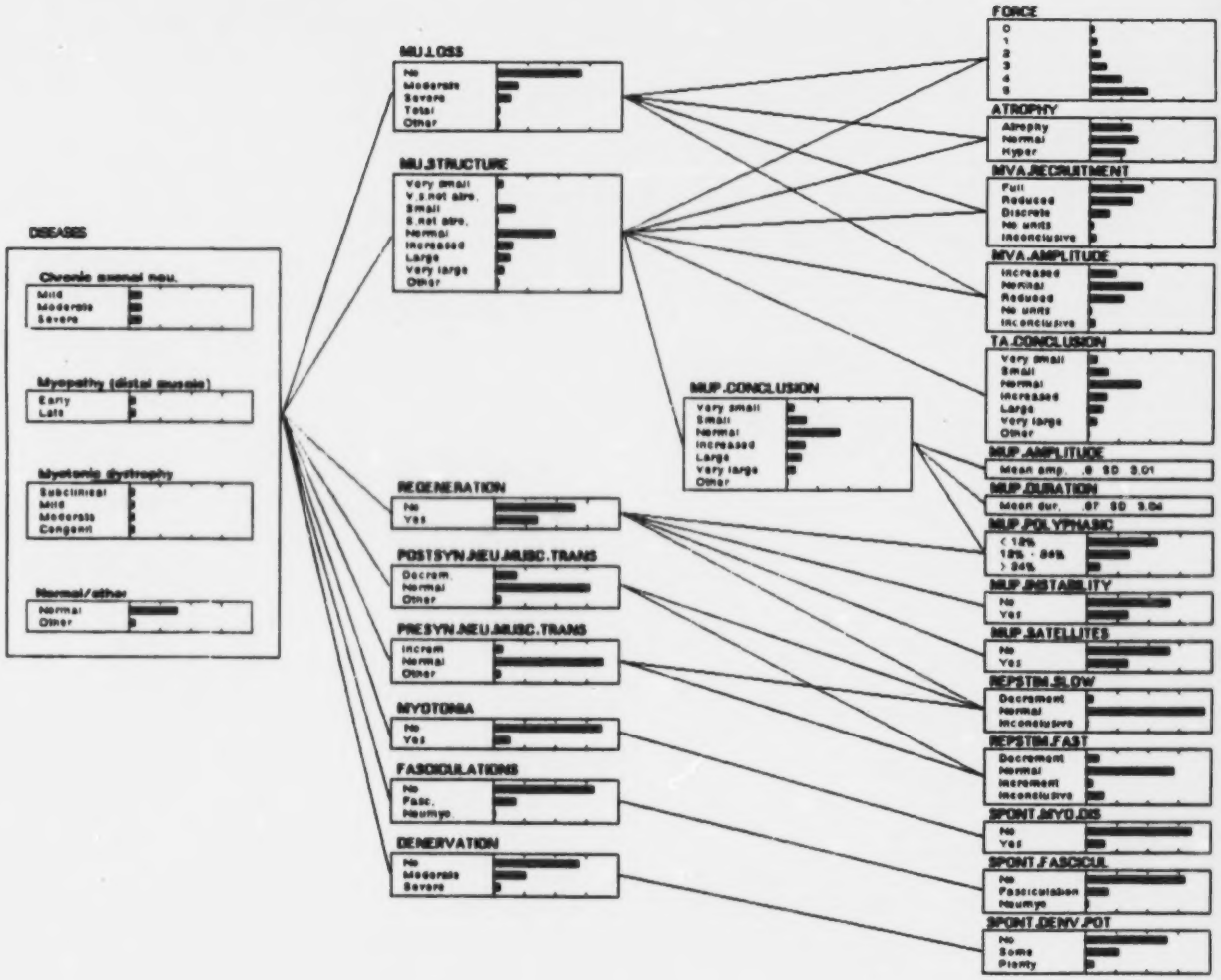


FIGURE 2

The MUNIN Network: Initial Status

(From Andreassen et al., 1987)

BEST COPY AVAILABLE

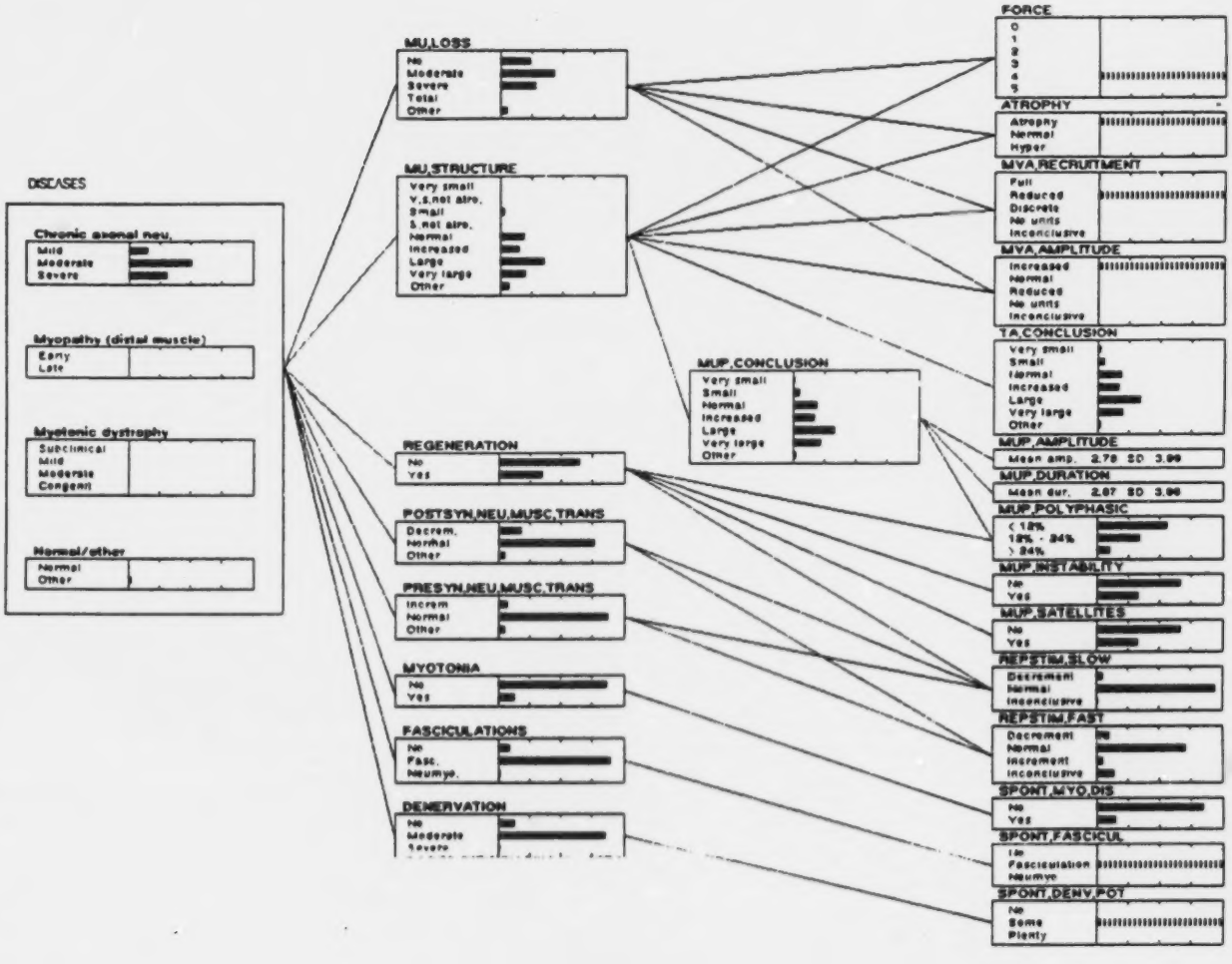


FIGURE 3

The MUNIN Network: After Selected Observations

(From Andreassen et al., 1987)

BEST COPY AVAILABLE

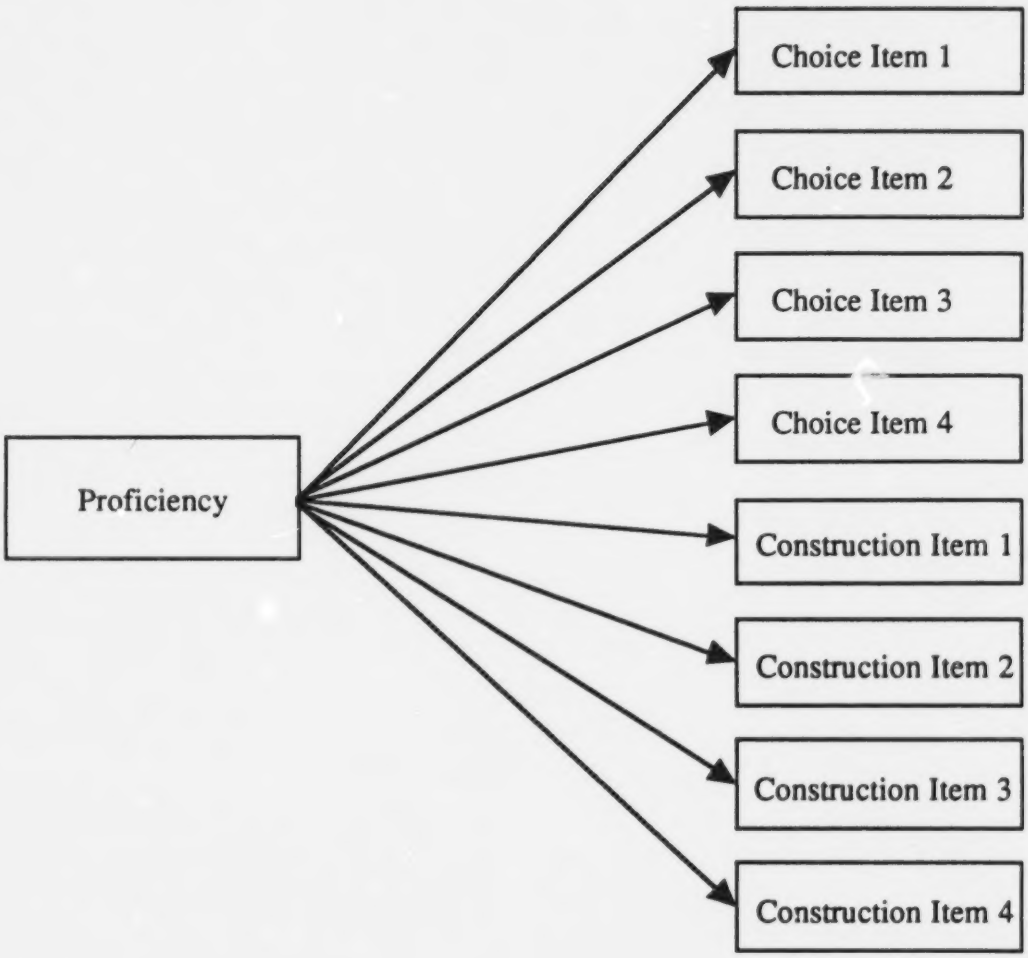


Figure 4

Network for a Single Proficiency, No Criterion Available

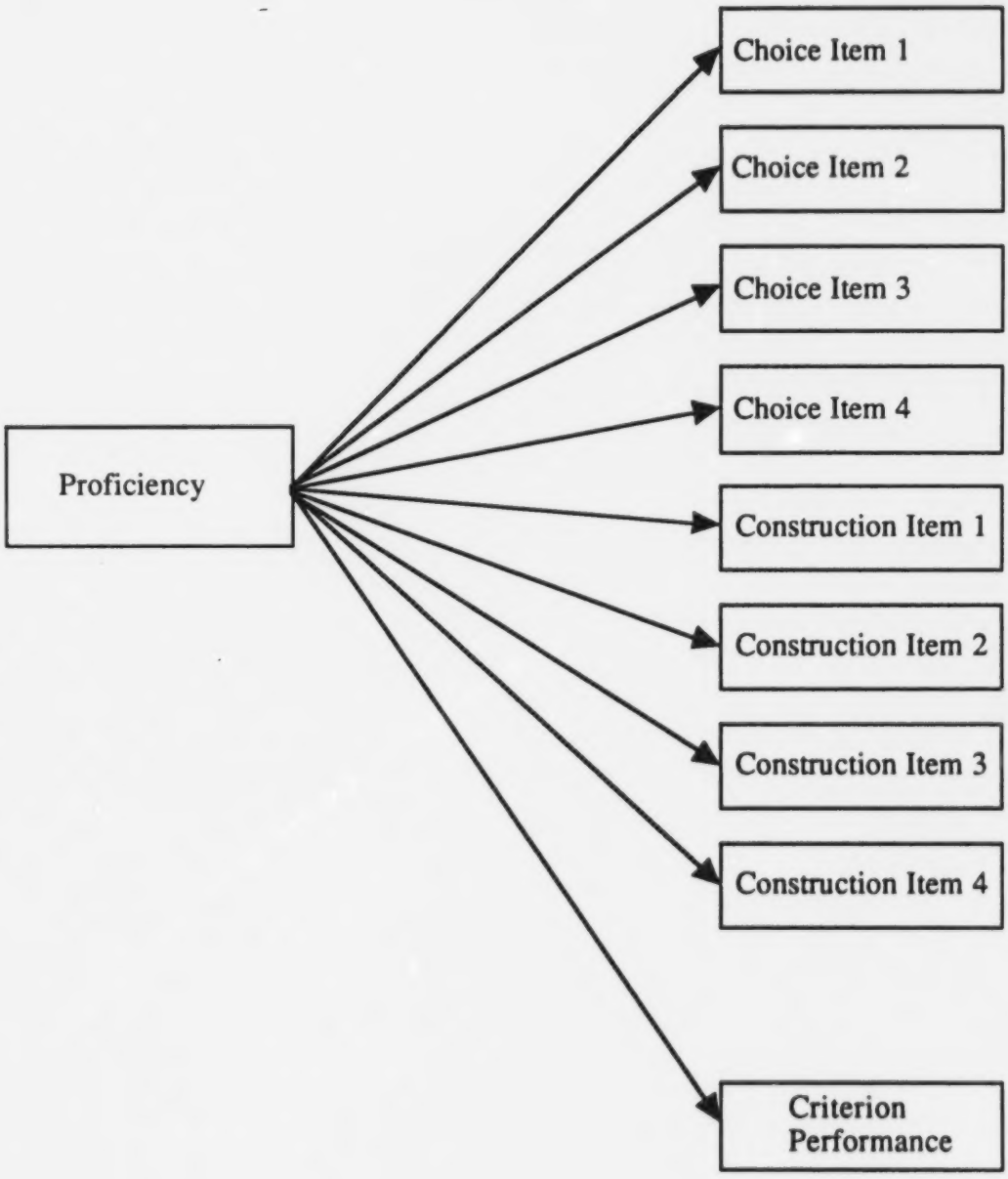


Figure 5

Network for a Single Proficiency, Criterion Available

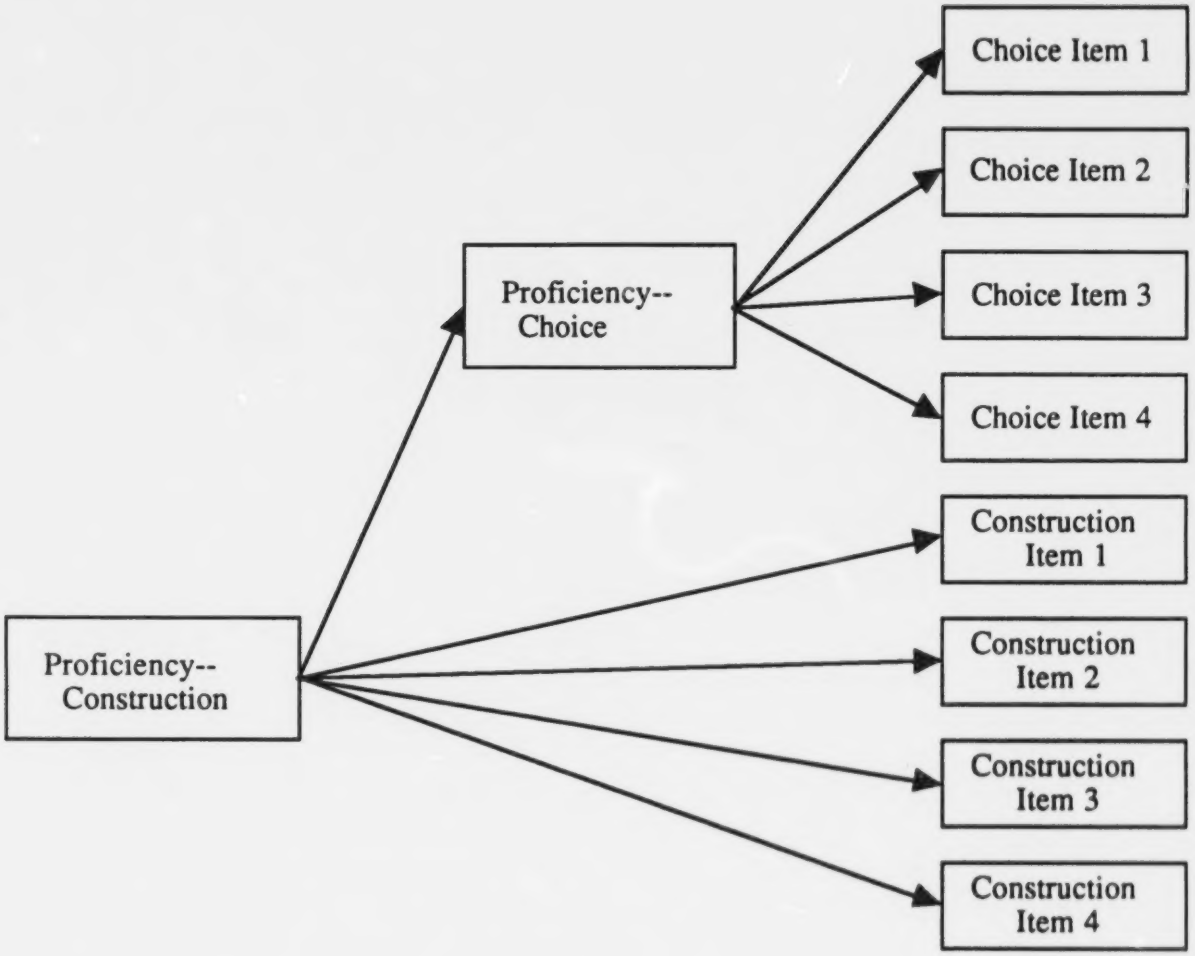


Figure 6

Network for Distinct Proficiencies, No Criterion Available

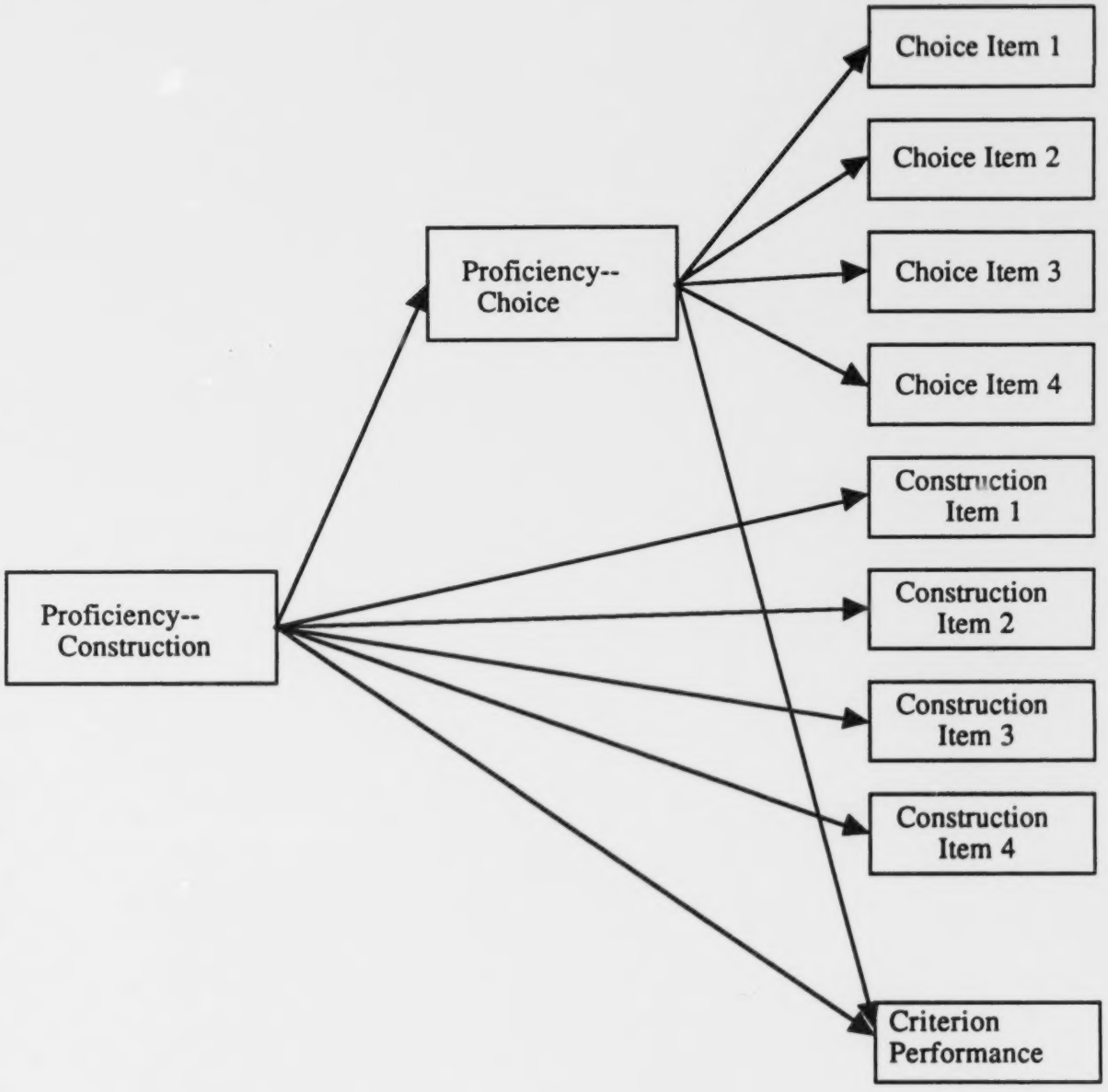


Figure 7

Network for Distinct Proficiencies, Criterion Available

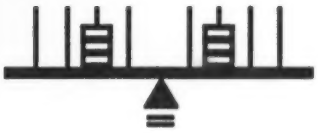
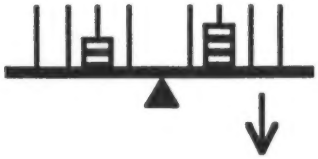
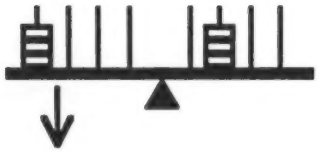
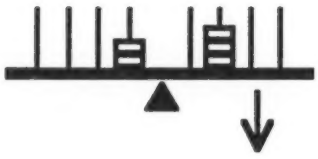
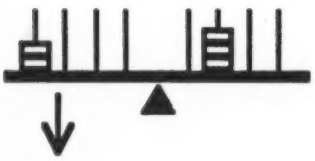
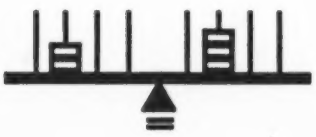
Item Type	Sample Item	Description
E		**Equal problems** (E), with matching weights and lengths on both sides.
D		**Dominant problems** (D), with unequal weights but equal lengths.
S		**Subordinate problems** (S), with unequal lengths but equal weights.
CD		**Conflict-dominant problems** (CD), in which one side has greater weight, the other has greater length, and the side with the heavier weight will go down.
CS		**Conflict-subordinate problems** (CS), in which one side has greater weight, the other has greater length, and the side with the greater length will go down.
CE		**Conflict-equal problems** (CE), in which one side has greater weight, the other has greater length, and the beam will balance.

Figure 8

Sample Balance Beam Items

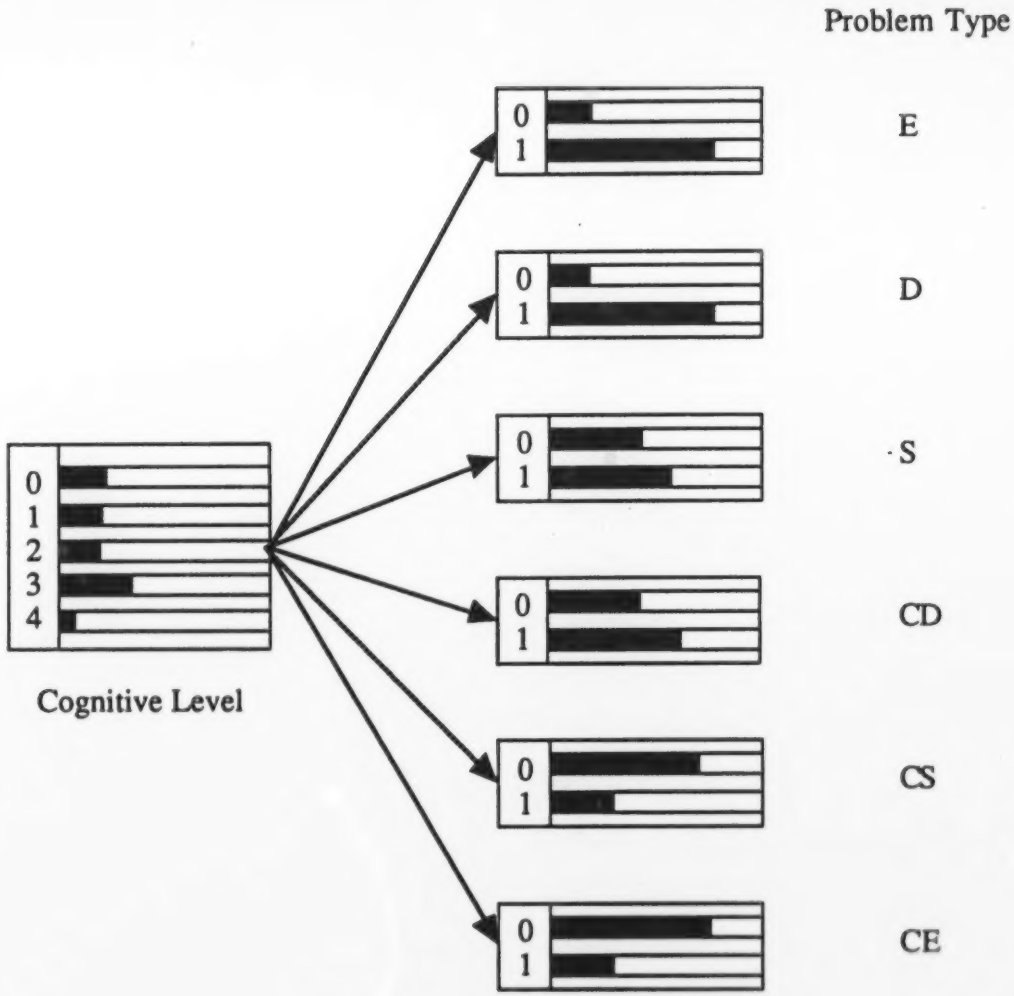


Figure 9

Initial State in an Inference Network for Balance Beam Examplar

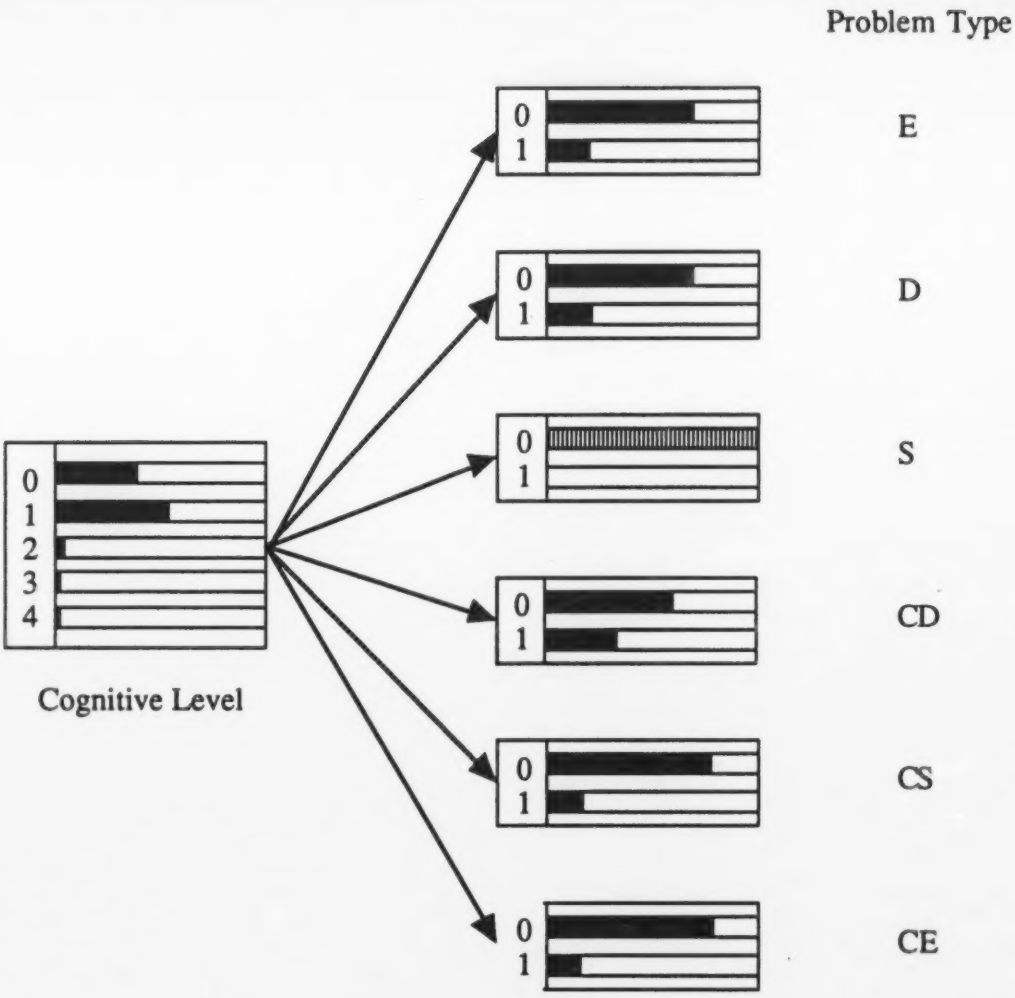


Figure 10

State of Knowledge after an Incorrect Response to an S Task

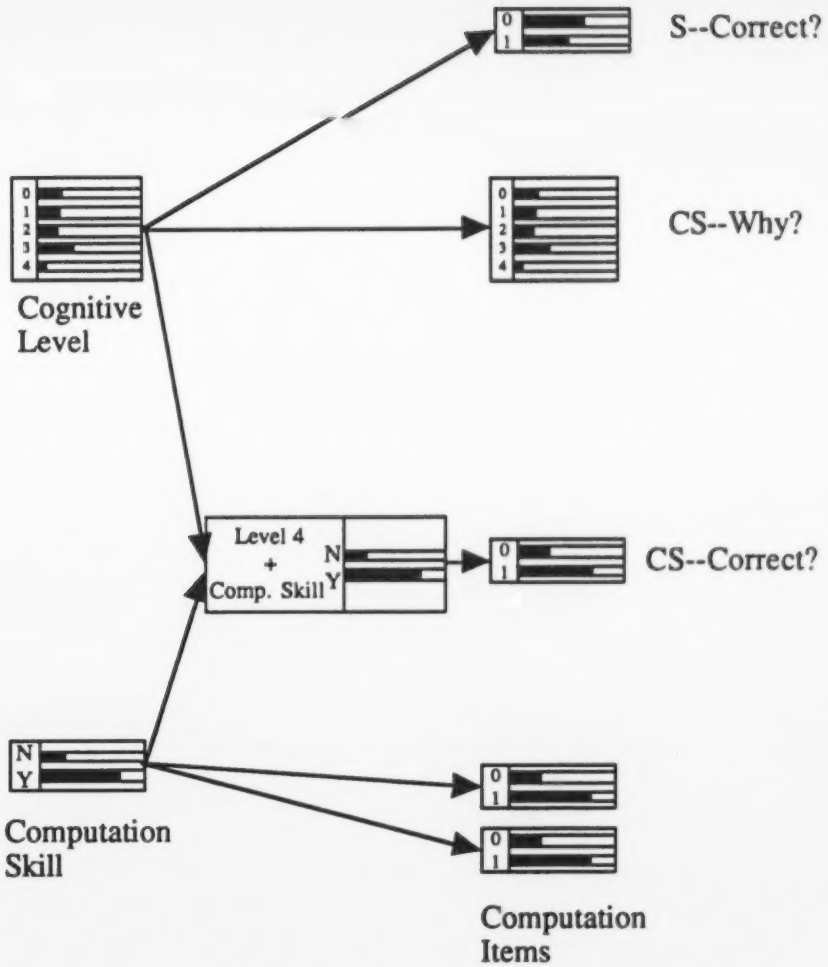


Figure 11

An Inference Network for an Extension of the Balance Beam Exemplar

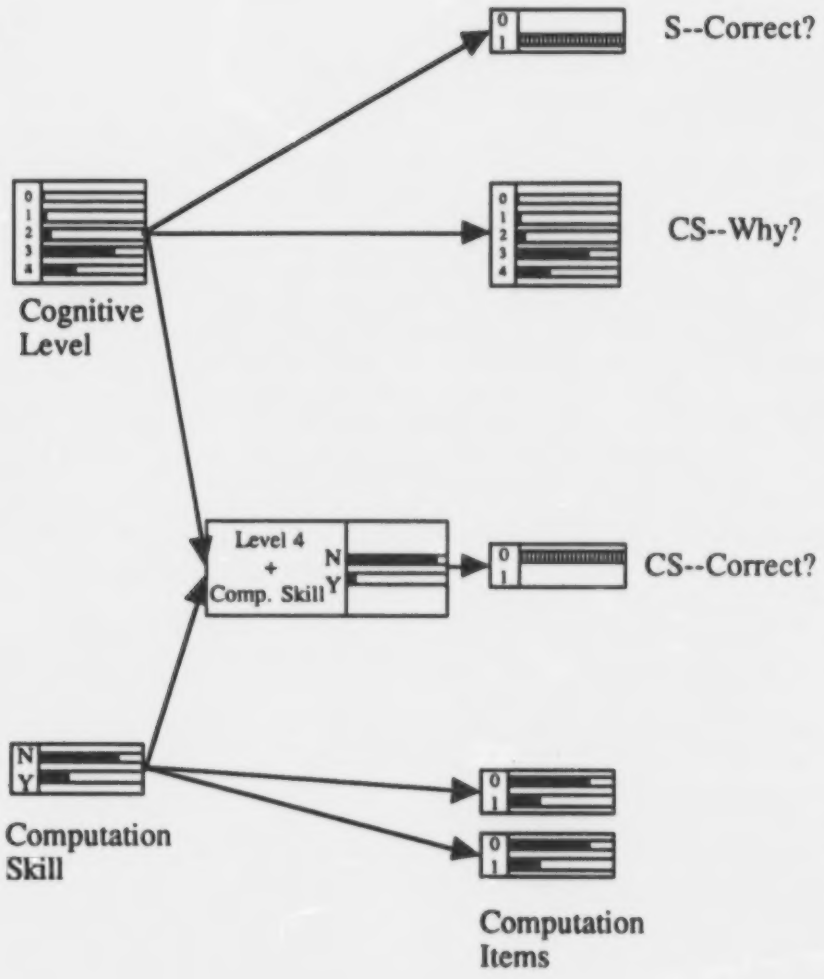


Figure 12

An Extended Inference Network: Knowledge after Two Responses

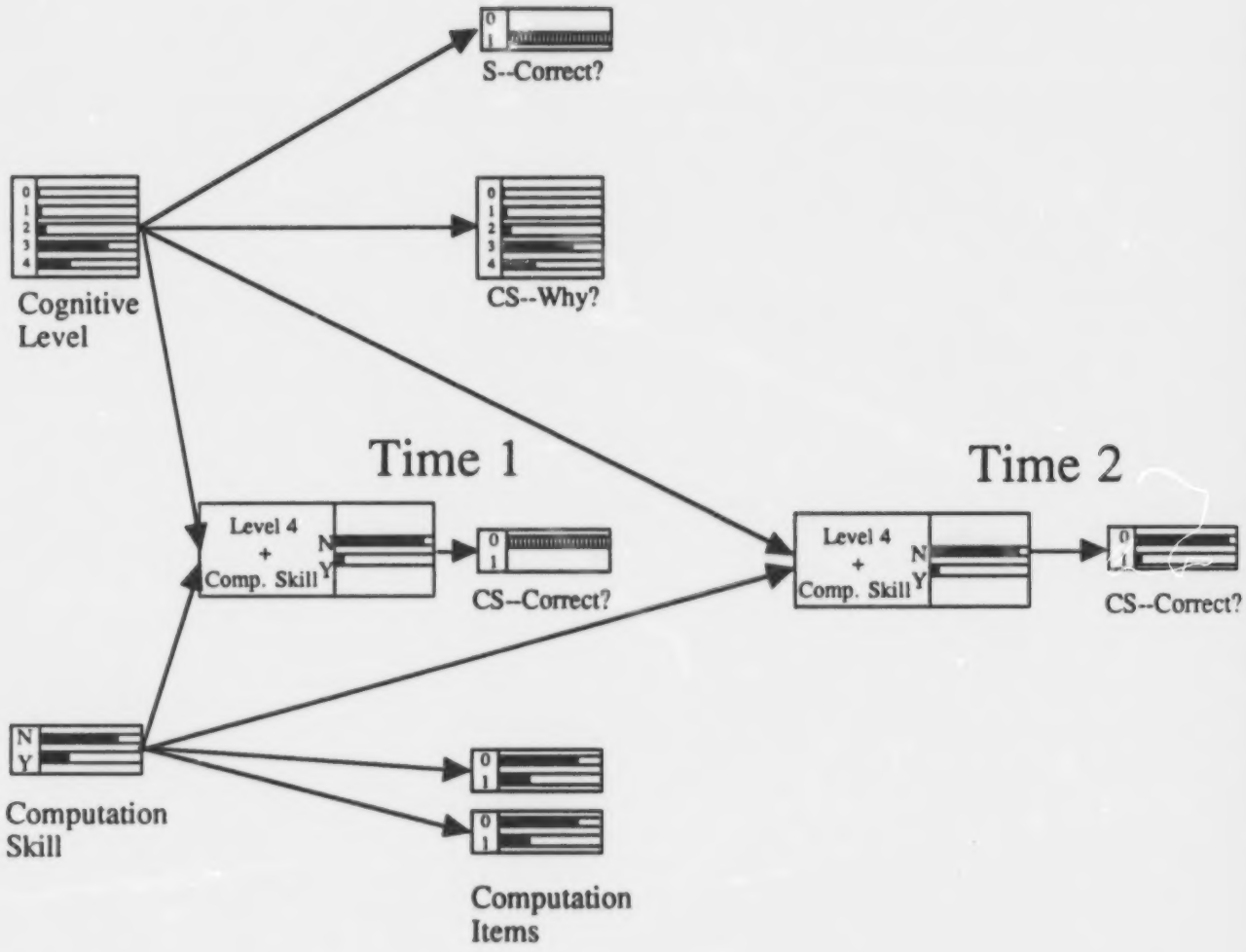


Figure 13

An Extended Inference Network: Prediction of CS at Time 2

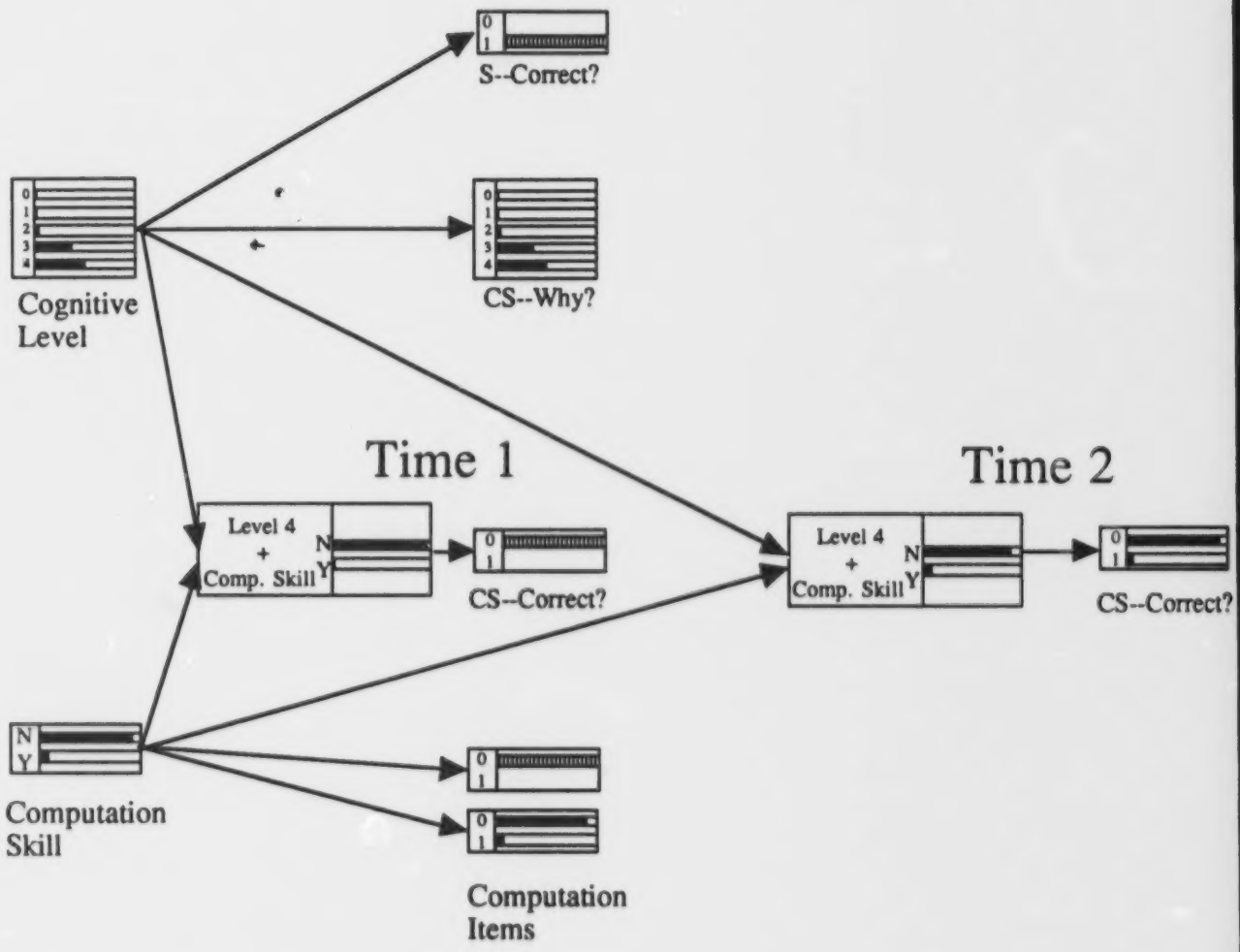


Figure 14

An Extended Inference Network: Knowledge after Three Responses, Including Computation

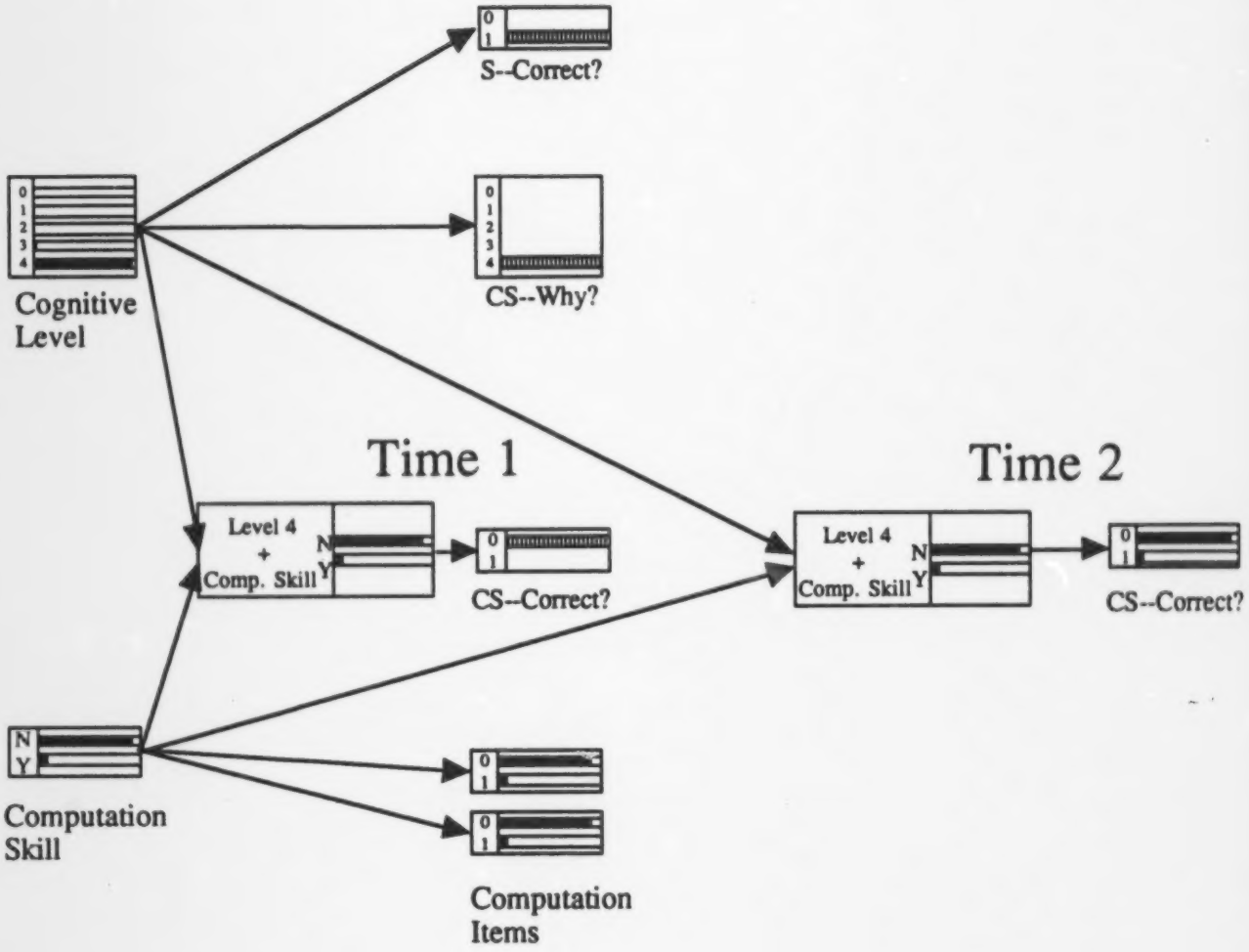


Figure 15

An Extended Inference Network: Knowledge after Three Responses, Including "Why?"

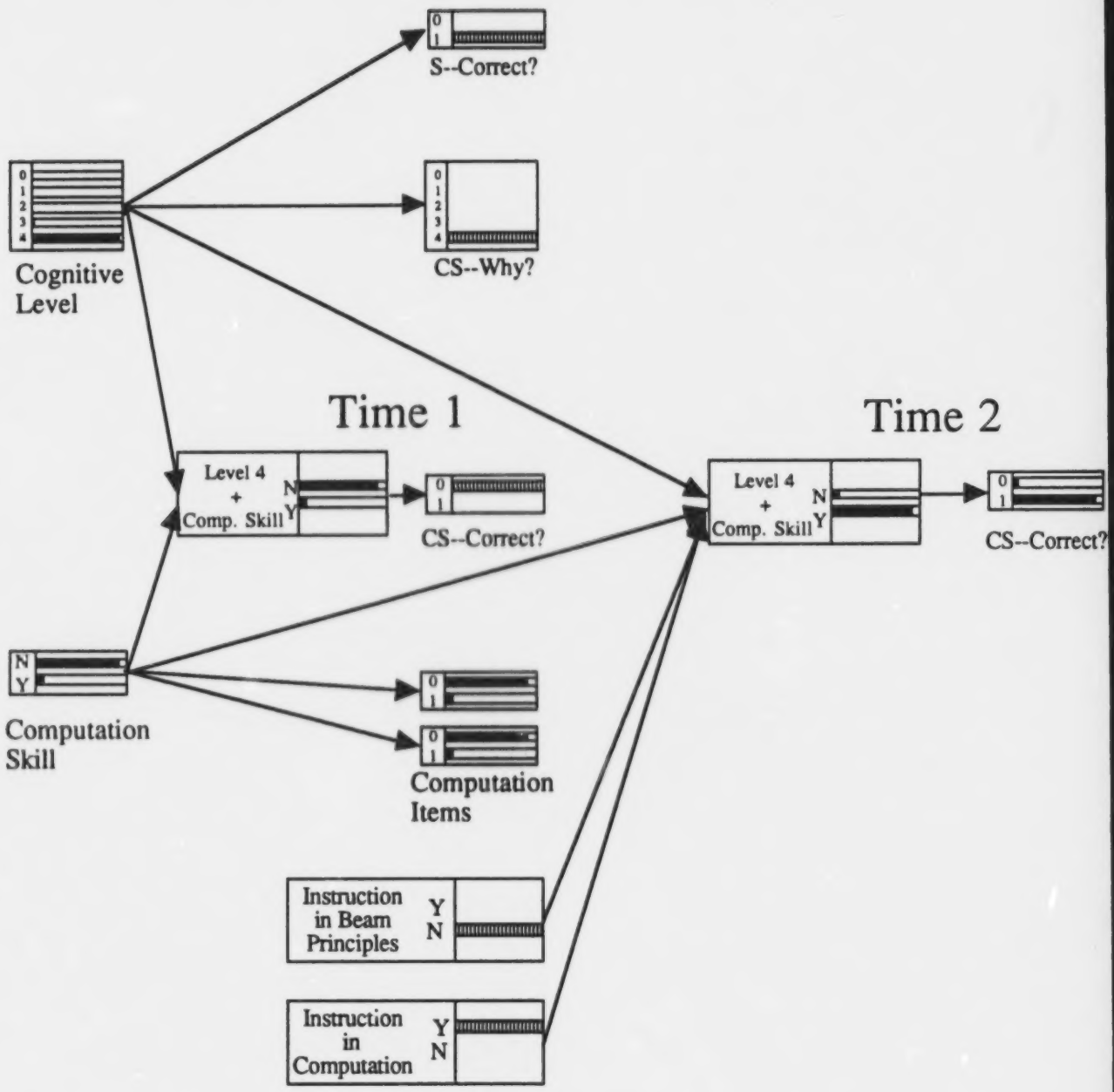


Figure 16

An Extended Inference Network: Instruction in Computation